DOCUMENT RESUME

ED 422 397 TM 028 959

AUTHOR Crehan, Kevin D.; Hudson, Rhoton
TITLE A Comparison of Two Scoring Strategies for Performance Assessments.
PUB DATE 1998-04-00
NOTE 10p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (San Diego, CA, April 14-16, 1998).
PUB TYPE Reports - Research (143) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Comparative Analysis; *Constructed Response; Cost Effectiveness; *Elementary School Students; Grade 5; Intermediate Grades; Models; *Performance Based Assessment; Reliability; *Scoring; *Writing Tests
IDENTIFIERS Scoring Rubrics

ABSTRACT

The aim of this study was to explore a method of improving the objectivity, reliability, and efficiency of scoring performance assessments that involve constructed written responses. Millman (1997) has suggested an alternative to using model responses at each score category. The proposed strategy, hypothesized to increase scorer reliability and cost effectiveness, would model answers judged to be halfway between the score categories. This paper reports on a small study designed to compare a scoring method using model responses at each category to a variation of Millman's suggested alternative. Existing student responses to a fifth grade reading prompt from a large school district's assessment program were used. Twenty volunteers (graduate students) served as raters, and 200 responses to the same prompt were divided into 5 groups of 40 responses. Two raters from each scoring group scored the same 40 papers, allowing the comparison of 2 scores for each response under each scoring condition. No differences were detected between the scoring methods. This may be due to the difficulty of obtaining agreement on borderline responses to be used in training, or it may represent the absence of a consensus on borderline anchor papers. In conclusion, it is stated that no evidence is found to differentiate levels of rater agreement between using judgments of dominance and judgments of proximity. Appendixes present two study scoring rubrics. (Contains one table and nine references.) (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 422 397

A Comparison of Two Scoring Strategies for Performance Assessments

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

Kevin Crehan

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

Kevin D. Crehan

University of Nevada, Las Vegas

Rhoton Hudson

Clark County School District

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

While much attention is currently being given to discussion of emergent conceptualizations of validity evidence, unresolved concerns remain for the more basic issues of objective and reliable scoring of performance assessments, especially for writing products. The focus on this study is on exploring a method of improving objectivity/reliability and efficiency of scoring performance assessments which involve constructed written responses.

Moss (1992) and Linn (1993) observed that there is a problem concerning comparability of scores assigned by different raters. This source of error is attributed to the necessity of reliance on professional judgment in scoring performance assessments. However, Linn notes that, with careful training of raters on well-designed rubrics, the error variance due to raters is less than that due to task specificity. Linn reports satisfactory generalizability across raters has been observed in a number of contexts, given explicit scoring rubrics with intensive reinforced training. Additionally, the California Assessment Program has established an inter-rater reliability of .90 for their writing assessment by using procedures

Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego, CA, April, 1998

TM028959

which include providing sample anchor papers for each rater and recirculating previously scored papers to check on stability (U.S. Congress, Office of Technology Assessment, 1992). Shavelson, Baxter, and Pine (1992) observed the reliability and validity of performance assessments in the 5th and 6th grade science curriculum. They asked the question: How large a sample of observers is needed to produce reliable measurement? Their results found inter-rater reliability to be consistently high in evaluating student performance on complex tasks, high enough to conclude that a single rater provides a reliable score.

While the reports of Linn (1993) and Shavelson et al. (1992) are promising, earlier writers are less encouraging. In reviewing the pros and cons of essay examinations, Coffman (1971) reports a lack of conformity in scoring among different raters. Coffman and Kurfman (1968) found two raters differing by 142 points on a set of 60 papers, which suggests that, if a specific score is needed to pass an examination, then the severity of the person scoring the paper will determine whether it passes or fails. Coffman also found that raters can vary in how they distribute grades across the score scale and in the value they place on different papers as well as in how strictly they score. In his review, Coffman observed inter-rater reliability coefficients ranging from .35 to .98, dependi ng on the context, content, or number of raters scoring. Godshalk, Swineford, and Coffman (1966) found that essay examinations read toward the end of a several day scoring session tend to receive lower scores than those read earlier in the grading session. Training included rating sample papers and comparing scores with scores given by other raters. For a large field test, the inter-rater reliability was only .672 for three readers. Crehan, Hudson, and Costa (1994) also observed marginal inter-rater agreement in scoring writing performance assessments.

Low rater agreement in this study may have been due, in part, to the variability of responses among examinees. Millman (1997) would agree that the problem of scoring objectivity is probably highest when the examinee is given some freedom in responding, as often is the case in the assessment of writing ability. Typically, a form of analytical or holistic scoring is employed in these instances since an unanticipated range of responses may demonstrate similar writing ability. Under these scoring schemes, the rater is trained on model responses at each score level and the rating task is to assign each writing product to a score category. Since the variety of responses which could be generated at each level of writing skill is large and the number of model responses is small, the task of rating is difficult.

Millman (1997) suggests an alternative to using model responses at each score category which he hypothesized will increase scorer reliability and cost effectiveness. The proposed strategy would model answers judged to be halfway between the score categories. The scoring task would then be to rate responses as better or worse than the model response. Millman predicts that the "judgments of dominance will be more reliable than judgments of proximity. (p.13)" This is a small study designed to compare a scoring method using model responses at each score category to a variation of Millman's suggested alternative.

Methods

Existing student outcomes to a fifth grade "response-to-reading" prompt from the assessment program of a large school district were used in this study. The district holistic scoring rubric (see Appendix A) was modified from describing a response appropriate for a given score category to one which suggested borders between score categories (see Appendix B). The attempt to identify a sufficient number of consensus anchor papers between score

categories was not successful and it was decided to use a range of responses for each score point as anchor and training papers. Twenty volunteers, ten from each of two graduate research methods classes, served as raters for the study. On consecutive days, an experienced scoring trainer gave each group of ten raters one and one-half hours of training in their assigned scoring method using the same eight anchor and eight training practice papers using the appropriate rubric for each condition. Two hundred responses for the same fifth grade prompt were divided into five groups of forty responses. Two raters from each scoring group scored the same forty papers, allowing the comparison of two scores for each response under each scoring condition.

Results

Table 1 reports percents for same score, agreement within one score category, and agreement within two score categories, generalizability coefficients, and scoring time for the two scoring methods. No differences were detected on any of these indices.

Discussion

The failure to find any differences between the scoring methods may be due to the difficulty of obtaining agreement on borderline anchor responses to be used in training. Or perhaps the absence of a difference explains the inability to reach consensus on borderline anchor papers. In any event, not having consensus borderline anchor papers prevented a good test of Millman's (1997) suggested scoring variation. Except for the difference in emphasis during training, the scoring conditions were too similar.

The score categories each contain a range of performance and, considering the degree to which rater judgment is involved, the boundaries are fuzzy at best. In retrospect

(regrettably), if consensus were reached on borderline responses, this consensus would have defined another score category.

In conclusion, no evidence was found to differenciate the levels of rater agreement between using judgments of dominance and using judgments of proximity.

REFERENCES

Coffman, W.E. & Kurfman, D.A. (1968). A comparison of two methods of reading essay examinations. American Educational Research Journal,5,99-107.

Coffman, W.E. (1971). Essay examinations. In R.L. Thorndike (Ed.) Educational Measurement (2nd Ed.). Washington, D.C.: American Council on Education.

Crehan, K.D., Hudson, R., & Costa, J.S. (1994). Introducing locally developed performance measures into a school assessment program. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Godshalk, F.I., Swineford, F., & Coffman, W.E. (1966). The Measurement of Writing Ability. New York: College Entrance Examination Board.

Linn, R. L. (1993). Educational assessment: Expanded expectations and challenges. Educational Evaluation and Policy Analysis,15,1-16.

Millman, J. (1997). Ideas for thesis and other research in educational measurement and related topics. Invited address presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.

Moss, P. A. (1992). Shifting conceptions of validity in educational measurement: Implications for performance assessment. Review of Educational Research,62,229-258.

Shavelson, R. J., Baxter, G. P., & Pine J. (1992). Performance assessments: Political rhetoric and measurement reality. Educational Researcher,21,22-27.

U.S. Congress, Office of Technology Assessment. (1992) Testing in American Schools: Asking the Right Questions (OTA-SET-519), Washington, D.C.: U.S. Government Printing Office.

TABLE 1

RATER AGREEMENT, GENERALIZABILITYCOEFFICIENTS,

AND AVERAGE SCORING TIME FOR THE TWO SCORING METHODS

	PROXIMAL SCORING	DOMINANCE SCORING
N OF RATERS	10	10
RESPONSES RATED	40	40
PERCENT SAME RATING	44	42
PERCENT WITHIN ONE	49	50
PERCENT WITHIN TWO	7	8
GENERALIZABILITY	.75	.74
AVG SCORING TIME (MIN.)	58	62

Appendix A

THE FUN THEY HAD
STORY RUBRIC

Score four (4) if the student accurately and completely summarizes (not copies) the setting, the main characters, and the main events.

- Includes at least one detail about each "school"
- Events are related in correct order
- Events are stated explicitly rather than inferred through indirect language

Score three (3) if the student summarizes the setting, the main characters, and the main events with minor inaccuracies

- one detail about "school" is stated
- events not in correct order
- one event inferred

Score two (2) if the student summarizes the setting, the main characters, and most of the main events

- may contain major flaws in the story line
- may include irrelevant details
- may include some copying
- irrelevancies may detract from the story
- may generalize the characters
- one or more thing may be missing

Score one (1) if the student does not adequately summarize the setting, the main characters, and the main events

- may be substantially copied
- may be a retelling of the whole story
- setting may be unclear

Score zero (0) for no response or an inappropriate response

Appendix B

THE FUN THEY HAD STORY RUBRIC

The student accurately and completely summarizes (not copies) the setting, the main characters, and the main events.
- Includes at least one detail about each "school"
- Events are related in correct order
- Events are stated explicitly rather than inferred through indirect language

If the above is satisfied, award a score of four (4), if not ...

Summarizes the setting, the main characters, and the main events with minor inaccuracies
- one detail about "school" is stated
- events not in correct order
- one event inferred

If the above is satisfied, award a score of three (3), if not ...

Summarizes the setting, the main characters, and most of the main events
- may contain major flaws in the story line
- may include irrelevant details
- may include some copying
- irrelevancies may detract from the story
- may generalize the characters
- one or more thing may be missing

If the above is satisfied, award a score of two (2), if not ...

Does not adequately summarize the setting, the main characters, and the main events
- may be substantially copied
- may be a retelling of the whole story
- setting may be unclear

If the above is satisfied, award a score of (1), if not ...

No response or response in inappropriate - score zero (0)